Rate-Distortion Approach to Databases: Storage and Content-based Retrieval

Ertem Tuncel, Prashant Koulgi, and Kenneth Rose[1]
Dept of ECE, University of California Santa Barbara, CA 93106, e-mail: {ertem,prashant,rose}@ece.ucsb.edu

The problem of *similarity search* is central to a wide range of applications in multimedia databases. The degree of similarity is often quantified by a distance measure defined over the space of extracted *feature vectors*. In typical applications, since the volume of data is huge, and the feature vectors are of high dimensionality, it becomes necessary to access the feature vectors from a hard storage medium during the search. Since I/O operations on hard storage devices are slow, the time complexity of the search is dominated by the I/O time.

One of the most effective approaches to reduce the time complexity is compromising the search accuracy by accessing *compressed* feature vectors (e.g., see [2]). The processing time is then approximately proportional to the rate R at which the sequence of feature vectors is compressed. We characterize the set of all achievable time-accuracy pairs in Theorem 1.

If the database is very large, the data must also be stored in compressed form, to reduce the *storage complexity*. Since the compressed feature vectors carry information about the corresponding data, their description could be embedded into the description of the data and be viewed as the base layer of a scalable coder. A high level diagram of the system is depicted in Figure 1. We study the trade-off between the first layer rate R_1, and the total rate R_2, employed to achieve distortions D_1 and D_2. Here, D_1 reflects the *accuracy of the search*, while D_2 measures the *data reconstruction quality*. Theorem 2 provides the characterization of all achievable (R_1, R_2, D_1, D_2). The conditions under which we do not suffer from a rate loss at any layer then follow as a corollary.

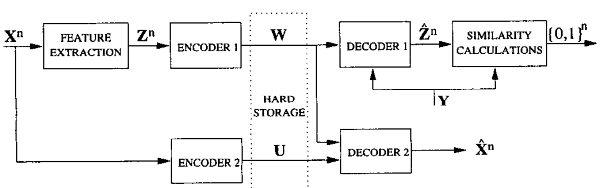

Figure 1: Block diagram of the system.

Let $\mathcal{X}$ and $\mathcal{Z}$ represent the data and the feature vector spaces, respectively. We denote by $\{X_i\}_{i=1}^{\infty}$ the data sequence, and by $Z_i \in \mathcal{Z}$ the feature vector extracted from X_i. Let $\hat{\mathcal{X}}$, $\hat{\mathcal{Z}}$ and $\mathcal{Y}$ denote the data reproduction, the feature vector reproduction, and the query spaces, respectively. In many cases of interest, $\mathcal{Y} = \mathcal{Z} = \hat{\mathcal{Z}}$, and $\mathcal{X} = \hat{\mathcal{X}}$. Finally, let $Y \in \mathcal{Y}$ denote the query vector. Although X_i, $\hat{X}_i$, Z_i, $\hat{Z}_i$, and Y are all vectors, we will consider them as "letters" of the corresponding super-alphabets. We assume that X_i are i.i.d. $\sim P_X(x)$, and that $Y \sim P_Y(y)$ is independent of $\{X_i\}_{i=1}^{\infty}$. The encoding of $\{X_i\}_{i=1}^{\infty}$ or $\{Z_i\}_{i=1}^{\infty}$ is performed after dividing the super-letter sequence into blocks of length n.

We introduce a query-dependent distortion measure $d_1' : \mathcal{Z} \times \hat{\mathcal{Z}} \times \mathcal{Y} \longrightarrow [0, \infty)$, in order to capture the dependency of the quality of quantization on the query point $y \in \mathcal{Y}$. Since feature extraction is a deterministic process, z is a deterministic function of x. Therefore we can equivalently consider $d_1 : \mathcal{X} \times \hat{\mathcal{Z}} \times \mathcal{Y} \longrightarrow [0, \infty)$. We extend this measure to blocks of length n as $d_1(x^n, \hat{z}^n, y) = \frac{1}{n}\sum_{i=1}^{n} d_1(x_i, \hat{z}_i, y)$. The data reproduction quality is evaluated by another distortion measure $d_2 : \mathcal{X} \times \hat{\mathcal{X}} \longrightarrow [0, \infty)$, and $d_2(x^n, \hat{x}^n) = \frac{1}{n}\sum_{i=1}^{n} d_2(x_i, \hat{x}_i)$.

Definition 1 A pair (R, D) is *achievable* if for all $\epsilon > 0$, there exists an encoding function $f : \mathcal{X}^n \longrightarrow \mathcal{M}$ and a decoding function $g : \mathcal{M} \times \mathcal{Y} \longrightarrow \hat{\mathcal{Z}}^n$ such that $|\mathcal{M}| \leq 2^{n(R+\epsilon)}$ and $E\{d_1(X^n, g(f(X^n), Y), Y)\} \leq D + \epsilon$.

This is almost exactly the Wyner-Ziv problem with a side-information-dependent distortion measure [3, 1]. However, the side information Y is a single random variable, *not* a sequence.

Theorem 1 (R, D) is achievable iff $R \geq R_s(D)$, where

$$R_s(D) = \min_{\substack{P_{U|X}(u|x),\, \phi(u,y) : \\ E\{d_1(X, \phi(U,Y), Y)\} \leq D}} I(X;U) \ . \tag{1}$$

The expected distortion above is to be computed assuming $P_{X,Y,U}(x, y, u) = P_X(x) P_Y(y) P_{U|X}(u|x)$. Note that this is precisely the Wyner-Ziv characterization [1], since $I(X;U) = I(X;U|Y)$ when $Y - X - U$ and Y is independent of X.

Definition 2 A quadruple (R_1, R_2, D_1, D_2) is *achievable* if for all $\epsilon > 0$, there exist encoding functions $f_1 : \mathcal{X}^n \longrightarrow \mathcal{M}_1$, $f_2 : \mathcal{X}^n \longrightarrow \mathcal{M}_2$, and decoding functions $g_1 : \mathcal{M}_1 \times \mathcal{Y} \longrightarrow \hat{\mathcal{Z}}^n$, $g_2 : \mathcal{M}_1 \times \mathcal{M}_2 \longrightarrow \hat{\mathcal{X}}^n$, such that $|\mathcal{M}_1| \leq 2^{n(R_1+\epsilon)}$, $|\mathcal{M}_1||\mathcal{M}_2| \leq 2^{n(R_2+\epsilon)}$, $E\{d_1(X^n, g_1(f_1(X^n), Y), Y)\} \leq D_1 + \epsilon$, and $E\{d_2(X^n, g_2(f_1(X^n), f_2(X^n)))\} \leq D_2 + \epsilon$.

Theorem 2 A quadruple (R_1, R_2, D_1, D_2) is achievable iff there exist random variables $U \in \mathcal{U}$, and $\hat{X} \in \hat{\mathcal{X}}$, and a deterministic function $\phi : \mathcal{U} \times \mathcal{Y} \longrightarrow \hat{\mathcal{Z}}$, such that $P_{X,Y,U,\hat{X}}(x, y, u, \hat{x}) = P_X(x) P_Y(y) P_{U,\hat{X}|X}(u, \hat{x}|x)$, and

$$I(X;U) \leq R_1 \ , \tag{2}$$
$$I(X;U,\hat{X}) \leq R_2 \ , \tag{3}$$
$$E\{d_1(X, \phi(U,Y), Y)\} \leq D_1 \ , \tag{4}$$
$$E\{d_2(X, \hat{X})\} \leq D_2 \ . \tag{5}$$

Corollary 1 (Successive Refinability) The quadruple $(R_s(D_1), R(D_2), D_1, D_2)$ is achievable iff there exists $P_{U,\hat{X}|X}(u, \hat{x}|x)$ and $\phi(u, y)$ such that $X - \hat{X} - U$, and

$$I(X;U) = R_s(D_1) \ , \tag{6}$$
$$I(X;\hat{X}) = R(D_2) \ , \tag{7}$$
$$E\{d_1(X, \phi(U,Y), Y)\} \leq D_1 \ , \tag{8}$$
$$E\{d_2(X, \hat{X})\} \leq D_2 \ . \tag{9}$$

References

[1] T. Linder, R. Zamir, and K. Zeger. On source coding with side-information-dependent distortion measures. *IEEE Trans. on Information Theory*, 46(7):2697–2704, November 2000.

[2] R. Weber and K. Bohm. Trading quality for time with nearest-neighbor search. In *Int. Conf. on Extending Database Technology*, pages 21–35, Konstanz, Germany, March 2000.

[3] A. D. Wyner and J. Ziv. The rate distortion function for source coding with side information at the receiver. *IEEE Trans. on Information Theory*, 22(1):1–11, January 1976.

[1]This work is supported in part by the NSF under grants no. EIA-9986057 and EIA-0080134, the University of California MICRO Program, Dolby Laboratories, Inc., Lucent Technologies, Inc., Mindspeed Technologies, Inc., and Qualcomm, Inc.

0-7803-7501-7/02/$17.00©2002 IEEE.